

Pourquoi je ne suis pas encore passé à Unicode

Stéphane Bortzmeyer

<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 29 Mars 2006. Dernière mise à jour le 9 Juin 2006

<http://www.bortzmeyer.org/pas-encore-utf8.html>

Je suis un grand fan d'Unicode, le jeu de caractères qui enterre tous les autres jeux de caractères en permettant d'encoder toutes les écritures du monde. Alors, pourquoi est-ce que je ne l'utilise pas moi-même sur mon poste ?

J'ai même écrit un cours en français sur Unicode <<http://www.afnic.fr/doc/formations/supports>>. Mais, sauf exceptions, je ne l'utilise pas. J'édite les pages de mon blog, comme celle-ci, en Latin-1. Mes fichiers textes, par exemple mes sources LaTeX ou DocBook, ou encore mes sources écrits en Python sont également en Latin-1. J'envoie du courrier en Latin-1. Parmi les exceptions, notons que les pages de mon blog sont publiées dans un encodage d'Unicode, UTF-8 et que je configure toujours mes bases de données <<http://www.bortzmeyer.org/postgresql-unicode.html>> PostgreSQL pour utiliser Unicode.

Pourquoi ne fais-je pas plus d'Unicode ? Bien sûr, mon choix de n'utiliser que du logiciel libre limite parfois mes possibilités, par exemple pour programmer en Java ou bien pour voir des gadgets qui bougent en Flash. Mais, pour Unicode, le logiciel libre n'est pas complètement désarmé et il existe même de nombreuses et excellentes documentations sur Unicode avec Unix comme "*Step by step introduction to switching your debian installation to utf-8 encoding*" <<http://melkor.dnp.fmph.uniba.sk/~garabik/debian-utf8/howto.html>> ou "*The Unicode HOWTO*" <<http://www.linux.org/docs/ldp/howto/Unicode-HOWTO.html>>, *DebFrUTF8* <<http://wiki.debian.org/DebFrUTF8>> (spécifique à Debian), "*Make your system use unicode/utf-8*" <http://gentoo-wiki.com/HOWTO_Make_your_system_use_unicode/utf-8> (spécifique à Gentoo) ou encore "*The Unicode HOWTO*" <<http://www.linux.org/docs/ldp/howto/Unicode-HOWTO.html>> ou enfin et surtout "*UTF-8 and Unicode FAQ for Unix/Linux*" <<http://www.cl.cam.ac.uk/~mgk25/unicode.html>>. (Les utilisateurs de FreeBSD peuvent regarder "*Unicode Support on FreeBSD*" <<http://opal.com/freebsd/unicode.html>>, encore très sommaire.)

Le problème est que tout n'est pas "Unicodisé". Les partisans d'Unicode tout de suite mettent en général en avant le fait que tel ou tel logiciel accepte Unicode. Mais, sur Unix, on n'utilise pas qu'un seul

logiciel mais toute une boîte à outils dont la combinaison permet des choses extraordinaires. Je ne travaille pas qu'avec OpenOffice (en fait, je fais plus souvent des documents avec LaTeX qu'avec OpenOffice, ce qui nécessite apparemment d'installer une extension et d'ajouter `\usepackage[utf8x]{inputenc}` dans le source) et il me faut donc attendre, avant de passer à Unicode, que pas mal d'outils aient été adaptés ou aient un équivalent.

Si mon éditeur favori, Emacs, se débrouille maintenant à peu près normalement (mais cela a pris du temps et il est bien plus lent en mode Unicode, ce qui est pénible pour les vieilles machines), et que je peux éditer les pages de Wikipédia en Unicode, si j'arrive à lancer un terminal (`LC_CTYPE=fr_FR.UTF-8 xterm -u8 -fn '-misc-fixed-medium-r-normal--14-130-75-75-c-70-iso10646-1'` où des programmes comme `cat` ou `more` l'affichent proprement, bien d'autres ne sont pas "Unicodisés".

Mais c'est surtout `grep` qui me manquerait : je suis habitué à taper `grep écart *.tex` et il me trouve tous les fichiers contenant ce mot. Si j'encodais tout en UTF-8, je n'aurai plus d'outil de recherche. (Marc Baudoin me fait remarquer que Perl gère parfaitement l'UTF-8 et peut donc être utilisé pour écrire rapidement un "grep-like" ; en voici un exemple (en ligne sur <http://www.bortzmeyer.org/files/ugrep.pl>) et autre (en ligne sur <http://www.bortzmeyer.org/files/ugrep.py>) en Python. Cela marche, mais c'est peu pratique d'avoir deux outils différents.)

De même, mon outil d'impression de fichiers textes, `a2ps`, ne comprend pas l'UTF-8. Je ne vais quand même pas lancer un monstre comme OpenOffice à chaque fois que je veux imprimer un script Python de dix lignes.

Bref, Unicode pour moi sera lorsque les applications traditionnelles auront été adaptées, pas juste deux ou trois gros logiciels.